30/1183US

UNITED STATES UTILITY PATENT APPLICATION

TITLE:

Computer Aided Ligand-Based and Receptor-Based Drug Design Utilizing Molecular Shape

FILED:

August 6, 2003

Inventor(s):

Randy J. Zauhar
William J. Welsh

Computer Aided Ligand-Based and Receptor-Based Drug Design Utilizing Molecular Shape

CROSS REFERENCE TO RELATED APPLICATIONS

This application claims priority to United States Provisional Patent Application Serial No. 60/401,637 filed August 6, 2002 the entire disclosure of which is herein incorporated by reference.

BACKGROUND

1. FIELD OF THE INVENTION

This disclosure relates to the field of computer aided drug design, and in particular, to utilizing shape similarity of molecules, as evidenced using ray-tracing techniques, to quickly compare compounds with each other and with receptors.

2. DESCRIPTION OF THE RELATED ART

It is generally presumed in medical science that a biologically-active (drug) compound having a particular effect on certain biological processes is biologically-active because of its ability to interact with a complementary receptor in the body. This interaction is further presumed to be due at least in part to the shape of the molecule and the electrostatic properties of the molecule. To put this in a simple fashion, a square peg will fit in a square hole while a round peg will fit in a round hole and drug design is simply locating the correctly shaped peg for the correctly shaped hole. Because of this apparent complementarity of the molecule to a receptor site, it is generally believed that by understanding the shape of a molecule which interacts with the receptor site, or the shape of the receptor site itself, molecules which have a shape complementary to the receptor site, or a shape similar to a known biologically-active molecule will also be complementary to the receptor, and therefore also biologically-active.

In ligand-based drug design, the underlying assumption is that a particular known biologically-active compound is complementary in shape to the desired target receptor and that the complementarity is responsible for the active effect. Therefore, in ligand-based design, the researcher attempts to locate other compounds having a similar shape to the known biologically-active compound. In receptor-based design, the structure of the target receptor is already generally known in atomic detail, and the goal is to identify compounds that would be complementary to the receptor both in shape and polarity. Both of these design methodologies therefore rely on an understanding of the shape of a particular compound, or the shape of a particular receptor, to locate new molecules having either a similar shape, or a complementary shape.

A number of methods have been devised for determining an indicator of the shape of a molecule or receptor, which can then be compared to other molecules of known shape to search for similarities or compliments. These methods have all met with limited degrees of success in their ability to accurately describe the shape of the molecules or receptors. Perhaps the most popular ligand-based strategy that takes shape explicitly into account is Comparative Molecular Field Analysis (CoMFA). In a CoMFA, the van der Waals and electrostatic fields of molecules are sampled over a grid superimposed on the molecule or receptor site. The values of these fields at the particular grid points are then used as descriptors in a regression model. CoMFA thus includes both molecular shape and polarity. While the CoMFA method can generally narrow a range of compounds to those which are generally similar, the CoMFA method has distinct problems in its application. In the first instance CoMFA uses a large amount of explicit information to encode shape, involving many grid points and geometric constraints. CoMFA also requires the inclusion of a overlaid gird on the molecule which generates questions of accuracy due to the effects of grid spacing and orientation of the molecules being compared.

Other ligand-based methods include the various methods for defining pharmacophore models. These represent ligand shape implicitly by incorporating some collection of hydrogen bond acceptors and donors and regions of steric bulk, and imposing inter-group distance constraints thereon. This 3D geometric information therefore provides an implicit representation of molecular shape but requires a large amount of implicit information to encode the shape.

Other approaches attempt to compute topological descriptors of molecules, beginning with chemical structure, or starting with the wave function, also often derive directly from the molecular shape. Further, methods based on chemical fingerprints generally also include implicit shape information, since only a restricted family of compounds will be compatible with the information contained in the fingerprint.

Receptor-based design strategies generally involve an explicit representation of shape derived from an atomic resolution structure of the active site. For example, UCSF DOCK methods pack the active site with spheres, producing an efficient representation of the volume available to accommodate a ligand, and combine this with positions of hydrogen bond acceptors and donors. Other docking algorithms such as FLOG, GOLD and FlexiDock use an all-atom representation of the active site, and thus represent its geometry in fine detail. Pharmacophore-like models can also be devised for receptors, and these include shape information in the same way as ligand-based models.

While all of these methods attempt to determine the shape of either an active molecule or a receptor site, the information is difficult to efficiently encode and to use in database searching. CoMFA and pharmacophore methods use a large amount of explicit information to encode shape, involving many grid points or geometric constraints. Further, scanning a chemical library with a pharmacophore query involves a significant amount of computation, since each molecule must be repositioned and flexed in order to determine if it can fit the model. Similarly, receptor-based strategies require many packed spheres and/or atom positions to encode the shape of the active site, and again scanning a chemical library with a docking program requires many detailed calculations for each compound considered. These calculations can involve molecular mechanics computations or, at the very least, so-called "bump checks" to test shape compatibility between receptor and ligand.

All of these computations take time. While improved efficiency in search methods has made approaches usable for screening existing chemical libraries, there is clearly no upper limit on the number of compounds that would like to be compared in the pursuit of new biologically-active compounds. Further, the faster, and more effectively, a database can be searched, generally the faster a new and useful compound can be discovered.

SUMMARY

Because of these and other problems described herein, and other issues known by those in the art, set forth herein are systems and methods related to the use of what are called "shape signatures" for compactly representing molecular shape, and how shape signatures can be used in both ligand-based and receptor-based molecular design. The systems and methods generally use ray-tracing to explore the volume interior to a ligand, or the space exterior to a receptor site. Shape signatures are then probability distributions derived from the ray-traces. Shape signatures can serve as compact descriptors of shape requiring modest storage space, and which may be quickly compared to test for shape similarity or complementarity.

Both the basic implementation of shape Signatures, which includes only shape information, as well as extensions which couple the existing ray-tracing technique with other computed molecular properties to define shape signatures with higher-dimensional domains include methods discussed herein. There is also illustrated an approach that includes the molecular electrostatic potential (MEP) to define a two-dimensional (2D) shape signature. These MEP-based 2D shape signatures combine shape and polarity information, and can be used to select molecules that are similar in shape and electrostatic potential to a query.

There is described herein, in an embodiment, a method for determining molecular shape comprising: having a first molecule with a determined solvent-accessible molecular surface which can be visualized by a processor; visualizing said molecular surface as a series of small area elements; said processor performing a ray-trace, said ray-trace comprising: starting a ray at a start point in one of said small area elements; propagating said ray in a first direction until said ray impacts said visualized molecular surface at an impact point; reflecting said ray from said visualized molecular surface as if said molecular surface was perfectly reflective of said ray;

recording characteristics of said ray; and repeating said steps of starting, propagating, reflecting, and recording using the impact point of the prior iteration as the start point of the next iteration until a stop condition from a set of stop conditions is reached; and computing a probability distribution based on characteristics of said ray, said probability distribution providing an indication of said molecule's shape.

In an embodiment the characteristics of said ray include at least one of: the distance of ray-trace segments and the points of impact and the stop condition may comprises the prior recording of a predetermined number of points of impact such as, but not limited to, 10,000, 50,000, or 250,0000. In an embodiment when said stop condition is reached, said method is repeated using at least one of a new starting point and a new direction for said ray.

In an embodiment, the step of recording includes recording said points of impact and determining and recording a molecular electrostatic potential (MEP) of said molecule at said points of impact and the probability distribution may comprise an MEP-based 2D shape signature.

In an embodiment, the probability distribution may comprises a 1D shape signature or a 2D shape signature.

BRIEF DESCRIPTION OF THE DRAWINGS

The patent or application file contains at least one drawing executed in color. Copies of this patent or patent application publication with color drawing(s) will be provided by the office upon request and payment of the necessary fee.

FIG. 1 provides a depiction of the structure of Indinavir, including the substituents used in Example 4.

FIG. 2 provides a depiction of the solvent accessible surface of Indinavir which has been triangulated using SMART.

FIG. 3 provides a depiction of the geometry used in a ray-trace. The component of the incoming ray parallel to the plane is unchanged by reflection, while the component perpendicular to the plane is reversed.

FIG. 4 provides for two different ray traces for the surface in FIG. 2. Fig. 4A has only 100 reflections while FIG. 4B has 10,000 reflections.

FIG. 5 1D shape signatures based on ray-trace segment lengths for Indinavir. FIG. 5A has 10,000 reflections, FIG. 5B has 50,000 reflections.

FIG. 6 provides for a representation of an embodiment defining a two-dimensional (2D) shape signature.

FIG. 7 shows MEP-based 2D shape signatures for Indinavir. Fig. 7A uses 10,000 reflections, FIG. 7B uses 50,000 reflections.

FIG. 8 shows an example of a ray-trace including several segments that span the diameter of a single atom.

FIG. 9 provides a flowchart for an embodiment of a process for obtaining a shape signature for a molecule.

FIG 10 provides for score distributions for Example 3 using a logarithmic axis. FIG. 10A is for a 1D shape signature, FIG. 10B is for a 2D shape signature. Both utilize the L_l metric.

FIG.11. provides for score distributions for Example 3 using a linear axis. FIG. 11A is for a 1D shape signature, FIG. 11B is for a 2D shape signature. Both utilize the L_l metric.

FIG. 12 provides a simplified flowchart of the steps performed by the ALMS computer program to attach hits to the framework.

FIG. 13. provides for an embodiment of ray-traces in HIV Protease subsites shown in FIG. 1. Protein atoms involved in defining a site are orange, framework atoms are colored by atom type. All subsite atoms appear in capped-stick rendering. FIG. 13A correspondence to substituent (103). FIG. 13B corresponds to substituent (105). FIG. 13C corresponds to substituent (107).

FIG. 14 provides for results of 1D shape signature comparisons of Tripos fragments using the L_l metric. FIG. 14A is a textual presentation, FIG. 14B provides molecular drawings.

FIG. 15 provides textual results for MEP-based 2D shape signature comparisons of Tripos fragments using the L_l metric.

FIG. 16 provides for results of 1D shape signature comparisons of Tripos fragments against the NCI database using the L_l and R_l metrics. FIG. 16A is a textual presentation, FIG. 16B provides molecular drawings.

FIG. 17 provides for results of an MEP-based 2D shape signature comparisons of Tripos fragments against the NCI database using the L_l and R_l metrics. FIG. 17A is a textual presentation, FIG. 17B provides molecular drawings.

FIG. 18 provides an embodiment of the best NCI Derived fragments in molecular form for subsites R_2 (103), R_3 (105), and R_4 (107) as well as the CAS number of the source compound and the optimized energy. Indinavir is shown for comparison.

FIG. 19 provides an embodiment of a selection of inhibitors constructed using shape signatures and ALMS. Energies are after 1000 steps (max.) minimization with MaxiMin2, and correspond to a binding-energy estimate as described below.

DESCRIPTION OF PREFERRED EMBODIMENT(S)

For the purposes of discussion, the embodiments discussed below will principally be applied to shape signatures for the structure of Indinavir (which is shown in FIG. 1). One of ordinary skill in the art would understand that the systems and methods may be used to determine biologically-active compounds of any type and for use with any receptor, and the use of HIV Protease and Indinavir as exemplary embodiments should in no way be taken as a limitation on the invention.

One of ordinary skill in the art will also recognize that utilizing shape signatures to search a chemical compound library or other database of chemical compounds requires the database to include shape signature information. This disclosure will presume that the database has been updated with shape signature information so as to be useable to perform such a comparison. It is recognized that the methods used for determining a shape signature of any compound, could be used to determine a shape signature of any compound and therefore existing compounds and receptors with known characteristics can have those characteristics updated to include a shape signature.

In the shape signatures approach to molecular drug design, the shape of a molecule is assumed to coincide with its solvent-accessible molecular surface, which is determined as is known to those of ordinary skill in the art by the points of contact of a rolling spherical probe. As the shape signatures approach preferably uses a detailed representation of the surface, this molecular surface is preferably broken into small area elements. This is preferably accomplished through the use of the Smooth Molecular Surface Triangulator algorithm (SMART) as described in Zauhar, R.J. SMART: a solvent-accessible triangulated surface generator for molecular graphics and boundary element applications. *J. Comput.-Aided Mol. Des.* **1995,** *9*, 149-159, the entire disclosure of which is herein incorporated by reference. SMART partitions the molecular surface into regular triangular area elements, which are well-suited to the computations that follow. FIG. 2 provides for a representation of Indinavir including the triangulated surface (201) generated using this embodiment.

The definition of the solvent-accessible molecular surface depends upon the choices of atomic radii, solvent probe radius, and the density of element corners (vertices) to be generated. In an embodiment, the PARSE atomic radii as described in Sitkoff, D.; Sharp, K.A.; Honig, B. Accurate Calculation of Hydration Free Energies Using Macroscopic Solvent Models. *J. Phys. Chem.* **1994,** *98*, 1978-1988, the entire disclosure of which is herein incorporated by reference, is used. The radius for the solvent probe is selected to be 1.4 Å, and vertices are spaced approximately 0.5 Å apart. One of ordinary skill in the art would understand, however, that alternative values, models, and selections can be used in alternative embodiments of the invention.

The volume defined by the molecular surface is explored using a modified form of ray-tracing. Ray-tracing is a technique known to those of ordinary skill in the art for computer animation. Ray-tracing in traditional use tracks the paths of light rays that emanate from some number of pre-defined light sources, and which are then reflected by objects in a scene so as to accurately portray lighting and shadow effects in the final animated scene. In its full realization, ray-tracing takes into account the material properties of objects as well as the atmosphere through which the rays travel.

In an embodiment of the shape signature application, the ray tracing technique is altered so as to consider only perfect reflection as illustrated in FIG. 3 from the molecular surface eliminating any refracting or other imperfect reflection. Further, there are no actual light sources, but rather each light source and light ray is instead a hypothetical source with a source at a randomly-selected point on the molecular surface and a direction in a randomly selected direction. Effectively the "ray" is a line with a first predetermined source point and a first predetermined direction that is allowed to propagate by the rules of optical reflection when interacting with the molecular surface. Physically, the ray can be thought off as the path of a single photon reflecting on a perfectly reflective complicated surface.

In an embodiment, a ray is initiated at the midpoint of a triangular surface element of FIG. 2, which may be chosen at random or through a particular algorithm, with initial direction defined by selecting a second point at random or via a particular algorithm in a hemisphere centered at the midpoint of the planar element and directing the ray from the first point toward the second point. If the ray-trace is being generated for a ligand or other small molecule, then the hemisphere preferably lies on the interior side of the element as determined by the outward-facing surface normal (which is defined by the SMART algorithm discussed above). If the shape of a receptor site is being defined, then the hemisphere preferably lies on the outward side of the receptor site, and the initial ray propagation is directed toward the exterior of the molecule. One of ordinary skill in the art would understand why these choices are preferably selected. To simply describe, in conjunction with a molecule, the volume is similar to that of a peg (a "solid") whereas for a receptor the volume is similar to that of a hole (a "void").

When performing an exterior ray trace (such as for a receptor), the user preferably supplies a list of atoms that define the receptor site of interest, and only those surface elements that are close to the site atoms are involved in ray propagation, either as initiation points for new rays or as reflection points. Otherwise the ray trace could define an unusable shape signature.

Once a ray is initiated, it is propagated by the rules of optical reflection. That is the ray (301) effectively forms a beam of light started at the first point and directed in the direction of the second point, the ray (301) continues in a straight path until it hits the surface of the molecule, the ray (301) is then redirected as shown in FIG. 3. The ray (301) then continues again on a straight path until it hits the surface (201) of the molecule where it is again redirected as shown in FIG. 3. This travel and redirection is continued until a predetermined stop condition is obtained. Each time the ray (301) hits the molecular surface (201) the point of impact is written to a file.

In order to prevent an endless cycle of reflection, and to prevent an endless loop or infinite condition which provides no information, the propagation of the ray (301) is terminated by the occurrence of a stop condition. In an embodiment the stop condition is the occurrence of any of the following events: the number of reflections (hits) equals a preselected number set by the user; the propagating ray (301) makes a "glancing" contact with a surface element, or strikes too close to the boundary between two adjacent elements to be able to mathematically compute the reflection angle, or the ray strikes no surface element and heads out to infinity, a situation occurring in exterior ray-traces.

In the first of these situations, the algorithm is finished and the values for the points of impact are recorded. In the other two situations, the ray-trace may be restarted at a newly-chosen point and direction on the molecular surface, may be restarted at the same point but with a new direction, may be restarted at a new point with the same direction, or the information can be stored and combined with the information from additional rays. The idea behind the first three of these options is that the ray-trace provides a fixed number of hits by a single path, effectively following a path of total internal reflectance. FIG. 4A shows a first ray-trace for the structure developed in FIG. 2 where the ray (301) has been reflected of the surface (201) 100 times (100 hits). FIG. 4B shows a ray trace for the same structure but with 10,000 reflection points. As can clearly be seen the ray (301) has effectively "filled-in" the inner surface of the surface (201) being no longer even recognizable as a series of lines.

As should be clear from FIG. 4B, the ray-trace provides raw information about three-dimensional shape, particularly by supplying a series of points which are all located on the molecular surface. While the ray-trace can be useful in itself, in an embodiment of the invention, probability distributions are computed which characterize the ray-trace more compactly and that can then be used as more compact descriptors of shape (and, as discussed later, molecular polarity). "Shape signatures," as that term is used herein, therefore comprise these ray-trace derived probability distributions. For the purposes of this disclosure, the probability distributions (shape signatures) will be represented as histograms, but one of ordinary skill in the art would recognize that other representations (such as, but not limited to, wavelets) may be used in other embodiments, and in fact the abstract mathematical concept may be used without any representation at all in a still further embodiment.

For further purposes of terminology, a line segment of a ray that connects two successive reflection points is called a “ray-trace segment”. Perhaps the simplest shape signature is the distribution of the lengths (magnitudes) of these segments. This is called the one-dimensional (1D) shape signature to emphasize that the domain of the probability distribution (namely segment length) has one dimension. Fig. 5 shows the distribution of segment lengths for Indinavir. FIG. 5A shows the distribution derived from a 10,000 point ray trace and FIG. 5B shows the distribution from a 50,000 point ray-trace. It is observed that signatures converge rapidly with increasing number of reflections, and are apparently not sensitive to the initiation point of the ray-trace.

In addition to the 1D shape signature, shape signatures can also be defined with higher-dimensional domains (e.g. 2D shape signatures, 3D shape signatures etc.), which incorporate additional molecular descriptors. One approach to generating 2D shape signatures is to associate a surface property, measured at each reflection point, with the sum of the segment lengths on either side of the reflection. This is shown in FIG. 6 with the length of the first segment (501) summed with the length of the second segment (503) and included with a property of the point (505). One of ordinary skill in the art would see that these examples of 1D shape signatures and 2D shape signatures are merely exemplary and a shape signature can be based on any variable or combination of variables.

One particularly useful property of a particular reflection point is the molecular electrostatic potential (MEP) of the reflection point. This MEP may be obtained through a variety of methodologies, one of which is discussed later in the Examples. Fig. 7 shows MEP-based 2D shape signatures for Indinavir using 10,000 (FIG. 7A) and 50,000 (FIG. 7B) reflections. MEP-based 2D shape signatures are joint probability distributions for observing a sum of segment lengths together with a particular value of the electrostatic potential at a given reflection point. They thus simultaneously encode information concerning shape and polarity.

As should be clear from the figures so far discussed, shape signatures are independent of molecular orientation, and furthermore require no overlay of a grid on the molecule as they relate instead to internal volume distances and reflection points.

The above has gone into detail about how a shape signature can be obtained for any particular relevant molecule. Once shape signatures are obtained, they can be compared to rapidly test for shape similarity between multiple molecules, and shape complementarity between molecules and receptor sites, to perform ligand-based and receptor-based searches.

In an embodiment of the invention, shape signatures may be compared by measuring the distance between the associated histograms (shape signatures), using metrics that can be computed quickly. The first metric (equation 1) is based on the L_I norm commonly used to compare functions:

$$L_1 = \sum_i \left| H_i^1 - H_i^2 \right| \qquad (1)$$

In the use of equation 1, i ranges over the union of all the bins for histograms H^1 and H^2 (it is assumed throughout that the bins for any two histograms will have the same alignment, and so have a simple one-one correspondence, a more complex equation is used if this is not the case). It is assumed in this equation that the probability distributions are normalized, so that the sum of the histogram heights over all the bins is unity; then under the L_1 metric the maximum distance between two histograms is 2, in which case the histograms being compared have no common support (i.e. no bin positions where both functions simultaneously have non-zero height). The minimum distance between two histograms, under this and most other acceptable distance measures, is zero (corresponding to the case where the distributions being compared are identical). Therefore, the lower the distance measurement generally the more similar the two shape signatures.

Histogram shape signatures often feature a dominant peak around 3Å. This arises from ray-trace segments that "measure" small-scale as opposed to large-scale molecular shape. This can be by measuring a small atomic "leg" of the molecule or similar issues. See FIG. 8. In an attempt to amplify the sensitivity to overall molecular shape when making comparisons, the following modified metric (equation 2) may be used in another embodiment:

$$R_1 = \sum_i d_i \left| H_i^1 - H_i^2 \right| \tag{2}$$

This is called a ramp metric since it weights the *i*th term in the sum by a ramp function (the length d_i associated with the *i*th bins of the histograms).

The analogues of the preceding metrics for 2D shape signatures are equations 3 and 4:

$$L_1^{2D} = \sum_i \sum_j \left| H_{i,j}^1 - H_{i,j}^2 \right| \tag{3}$$

$$R_1^{2D} = \sum_i \sum_j d_i \left| H_{i,j}^1 - H_{i,j}^2 \right| \tag{4}$$

where index i varies over the length indices of the bins, and j varies over the second dimension (which in all the examples provided herein will be an electrostatic potential scale (MEP) as that is the choice in a preferred embodiment). The shape signatures approach allows comparison using either of these metrics with very little computing time. Most arithmetic in an embodiment will need to be performed on histograms with fewer than 50 bins. 2D shape signatures obviously require more arithmetic operations than 1D signatures to compare, but the computational expense is still limited and still shortened compared to most existing techniques.

It has already been noted that 1D shape signatures will often include a large peak at about 3Å segment length. Closer examination of the ray-traces from which the shape signatures are derived reveal a large number of segments that span atomic diameters. An example is shown in FIG. 8 where segments (601) essentially span the diameter of atom (605). These segments encode relatively little information about the overall shape of the molecule (simply defining the shape of an atom).

In another embodiment, the ray-tracing may be modified so that segments that involve reflections at the same atom are discarded. In an embodiment, additional segments are then generated as needed to match the user-specified total number of reflections (so as to keep numbers constant). This so-called segment culling helps to ensure that more of the surviving segments encode useful information about the overall shape of the molecule being considered but does entail added computational expense which may or may not be useful in any particular embodiment. Further, a culled shape signature will generally not be comparable to a non-called shape signature. Therefore, a particular embodiment will generally select whether to use culling or not depending on the intended use and desired comparison of the shape signature.

In use, the shape signatures may be used to identify molecules, molecular fragments, and receptor sites using a relatively simple and easily computable metric. The shape signatures of molecules with known biological effect can then be compared against the shape signatures of other molecules to attempt to find those likely to have a similar biological effect. Alternatively, the shape signature of a receptor may be compared to the shape signature of a multitude of molecules to find molecules which will hopefully bind to the desired receptor.

In an embodiment, shape signatures may be applied to molecular comparison. It is presumed there is an existing database of molecules, which may be saved in any format but is preferably saved in Tripos MOL2 format. From this database of molecules, the ray-tracing procedure may be performed to generate a database of 1D and 2D shape signatures for the molecules. The 2D shape signatures will preferably be MEP 2D shape signatures and the 1D shape signatures single ray-trace segment lengths as discussed above, but that is by no means required.

To be most useable, and provide for the simplest computation, the molecules in the database preferably have 3D atomic coordinates, and at least partial atomic charges. The atomic charges may be assigned by any method as know to one of ordinary skill in the art, but in the discussed embodiments the Gasteiger method described in: Gasteiger, J.; Marsili, M. Iterative partial equalization of orbital electronegativity —Rapid access to atomic charges. *Tetrahedron* **1980,** *36*, 3219-3288, the entire disclosure of which is herein incorporated by reference, is used.

The ray-tracing and shape signature assignment will generally be performed by a computer or other processor. This may be a standalone workstation, a portable machine, or a client or server located on an interconnected network such as an intranet, an extranet, or the Internet. Further, the term "processor" does not require a single processor as multiple processors may be used in series or in parallel and still comprise a processor. While the processor will preferably be a general purpose computer including software to perform the desired functions, the processor may also comprise dedicated hardware or software including all special purpose hardware and software. The software may be of any type utilizing any operating system and programmed in any language, but preferably comprises a C-shell script to perform operations of FIG. 9 on the database entry for each molecule.

First the processor generates a triangulated surface in step (901). This process may be implemented by the script using SMART (discussed above) implemented as a C program. In step (903) the ray-tracing is performed. Generally step (903) will be the most computation intensive step. In order to provide for sufficient speed, a fast algorithm using a grid acceleration method is employed in an embodiment of the invention. The algorithm generates a file with a pre-determined number of ray-trace segments which is stored in memory in step (905). In step (907) the histograms or other shape signature information are accumulated by reading the ray-trace file, summing the occurrences of segment lengths, computing the electrostatic potential, or any other steps needed to obtain the shape signatures. The generation of histograms will generally utilize a bin size specified by the user which may be selected by any means known to those of ordinary skill in the art.

To compute the electrostatic potential in step (907), the processor uses, in an embodiment, the partial atomic charges contained in the molecular structure file and equation 5:

$$\Phi(r_p) = \sum_j \frac{q_j}{|r_p - r_j|} \tag{5}$$

where $\Phi(r_p)$ is the molecular electrostatic potential (MEP) computed at reflection point r_p and the index j ranges over all the atoms, which have positions r_j (measured in Å) and partial charges q_j (measured in elementary charge units). The MEP values computed at each reflection point, along with the sum of the segment lengths that adjoin the reflection point, are used to accumulate a MEP-based 2D shape signature as illustrated in Fig. 6.

The resulting histograms are written to an file in step (909) (preferably an ASCII file) in a format that includes all pertinent information in an understandable form. In step (911), another script (which is preferably implemented in PERL) adds the histogram information generated to a new database (or alternatively as new entries in the existing database) creating a database including the shape signatures of the various molecules, fragments, and/or receptors.

In an embodiment, the query used to scan the database of shape signatures is itself a database of signatures which could refer to a single object. The query database could be generated from a set of small molecules, in which case the same procedure described above is employed, or a receptor site in a step prior to step (901). In the latter case, the procedure is similar, except that an exterior ray trace is performed (typically over a protein), and the user will specify a set of atoms that define the receptor site. In either case, comparison of the query and target database is effected using a processor, such as that described above, that compares each histogram in the query database against those in the target database using a metric such as one of the ones described above. The processor then reports a hit list file of the best n hits for each of the queries where the variable n is selected by the user or based on particular relevance determinations.

The above described metrics (equations 1-4) can readily compare receptor sites and ligands for shape complementarity, it is less straightforward to measure electrostatic complementarity. However, there are numerous approaches which may be used. One might, for example, simply reverse the sign of the electrostatic field for either query or target signature, and proceed using either of the existing metrics; one then finds an exact match only if query and target signature have exactly complementary shapes, and electrostatic fields that are equal and opposite. This is an extraordinarily stringent criterion, which would lead to poor scores for clearly useful matches, but can be useful in some situations. Alternatively, other approaches might be used as discussed in the examples.

It is preferable to have some criteria for assessing the significance of the hit list file produced in a shape signatures search. Scores for 1D and MEP 2D shape signature searches are generally not normally distributed, and it is inappropriate to use z-scores to test for the significance of hits. However, for the queries considered in the examples below, meaningful hits appear in the extreme tail of the distribution, and the typical score cutoffs lead to selection of a very small percentage of compounds from the database. While the distributions of scores are of intrinsic interest, as a matter of practice, the user of shape signatures would decide at the outset how many hits to retain in a given search which is a more important limiting factor. Evidence as to the range of scores likely to correspond to close matches is helpful in determining if the number of hits collected was appropriate, but not for much else. The most important observation concerning the distributions is that close matches between compounds are usually found far from the median.

A shape signature procedure may be less straightforward to use in receptor-based approaches. Active sites of a receptor may differ dramatically in shape, and a method for restricting the ray-trace to a region of interest may not always be immediately apparent and sites may need to be omitted because of computational issues. However, the approach provides for a generally faster computational comparison if such setup is performed.

The shape signatures approach generally does not take into account synthetic feasibility of actually creating the desired molecules or compounds focusing instead on shapes of even hypothetical structure. Fragments are attached to the framework as discussed in Example 4 with no regard to the existence of a synthetic route for preparing the derivative, and with no regard to the cost or availability of the necessary reagents.

As a whole, shape signatures is a comparatively easy technique to apply. Particularly for ligand-based applications, the method does not require extensive experience in constructing queries, adjustment of numerous parameters, or sophistication in the interpretation of results, which can be serious drawbacks with other methods. Moreover, it directly addresses those features of molecules, namely their shape and surface properties, which are believed most critical to determining their biological activity. In this determination, it is also more efficient than previously known methods in its use of computational time and resources.

EXAMPLES:

EXAMPLE 1:

The shape signatures method was applied to the Tripos fragment database, a diverse collection of small molecules including heterocycles, carbohydrates, amino acids and nucleotides, which is supplied as a standard component of the SYBYL molecular modeling package. This database was especially useful for initial tests given its small size and its incorporation of multiple representatives of each family of compound (ensuring that a given query from the database will usually have several potential matches). Very small fragments were removed from the database at the start, and also some perfectly linear molecules (e.g. allene) which were not handled well by the SMART surface algorithm. This left a total of 235 compounds. Dummy atoms were removed from the amino acids in the database and the resulting empty valences filled with hydrogens. The sidechains of Glutamic acid, Aspartic acid, Lysine and Arginine were modified to correspond to the ionized form. Gasteiger charges were assigned to all the compounds in the final set, and 1D shape signatures (utilizing ray-trace segment length) and MEP-based 2D shape signatures were generated using either 50,000 or 250,000 reflections in each ray-trace in separate computational experiments. The shape signatures were then assembled into databases.

Each resulting database was compared against itself (i.e. each compound in the database was used as a query and compared against all the remaining compounds). The L_1 and R_1 metrics

(Equations 1-4) were used in separate comparison. For each query, the ten best (lowest-scoring) hit compounds were retained. This self-comparison was carried out for both the 50,000- and 250,000-reflection databases, using both 1D shape signatures and 2D shape signatures, and with segment culling either enabled or disabled. Examination of the hit compounds in the context of their scores were used to propose score cutoffs to distinguish those matches likely to be interesting.

FIG. 14A shows hits found for a selection of query compounds, with 50,000 reflections per histogram using 1D shape signatures and the L_1 metric, and with segment culling enabled and disabled. FIG. 14B shows structures of the top five hits for each of the six queries, for the case of segment culling enabled.

It is seen that the 1D shape signatures selected compounds chemically or structurally similar to the query. This observation is amplified by examining all of the available data for the Tripos database. One compound of a class generally selects all compounds of the same class present in the database. A fatty acid (Laurate) generally selected all other fatty acids present in the database; a carbohydrate (α-Dglucopyranose) generally selected other carbohydrates; an amino acid (Lysine) generally selected other amino acids and so on. Dispensing with the segment-culling procedure affects the size of the scores slightly, usually making the distances between histograms a bit smaller, but clearly has little effect on the rank order of hits in this example. Switching to the R_1 metric changed the size of the scores, but had little impact on the rank order of hits for most of the queries.

MEP-based 2D shape signatures produced results similar to those of the 1D search, but with some changes in hit ranking that exhibit sensitivity to the electrostatic properties of query and target compounds, and with a much smaller number of meaningful hits. This is not surprising, given that the MEP-based 2D shape signature searches select simultaneously on the basis of shape and polarity, and thus are more stringent than 1D shape signature searches. Examples of this are seen in FIG. 15 for MEP-based 2D shape signatures compared under the L_1 metric. For example, where the query Lysine selected Methionine among the top five hits when only shape was considered (FIG.14A), the 2D shape signature search results do not include this compound. Initial experience with this metric suggests that meaningful 1D shape signature matches between query and target usually involve distances of less than 0.1 probability unit, while useful 2D shape signature hits are within 0.2 of the query.

The results found when comparing the fragment databases prepared using 250,000 reflections per compound were essentially identical to those discussed above, for all combinations of search type (1D or 2D) and metric (L_1 or R_1). This indicates that 50,000 reflections per compound assures adequate convergence, at least for molecules found in the Tripos fragment database.

CPU times for the Tripos fragment database self-comparison (235 queries, 55,225 comparisons) under the L_1 metric on a 1.5 GHz Pentium processor were 20.5 sec (1D search) and 53.7 sec (2D-MEP search). Timings under the R_1 metric were essentially identical. These total times correspond to approximately 370 μsec for a single 1D shape signature comparison, 970 μsec for an MEP-based 2D shape signature comparison.

EXAMPLE 2

The National Cancer Institute compound database as bundled with the SYBYL UNITY tools was used as a source of molecules for creation of a shape signatures database with 1D shape signatures based on ray-trace segment length and MEP-based 2D shape signatures. The starting database was screened for all compounds with molecular weight less than 800 Da, yielding

113,826 molecules. Gasteiger charges were computed for all of the molecules in the resulting working set. 1D shape signatures and MEP-based 2D shape signatures were computed for all the compounds, using a sixteen-processor Beowulf cluster. It should be pointed out that each processor was simply allotted a fraction of the molecules to be analyzed, and there was no need to employ the use of parallel code. 50,000 reflections were generated in the ray-trace for each compound, and segment culling was employed, as described above. Of the compounds processed, about 0.4% failed (in every case due to an error in molecular surface generation), yielding a total of 113,331 compounds in the NCI shape signatures database used in subsequent work. Preparation of the database consumed approximately 100 hours wall-clock time on a sixteen-processor cluster of 450 MHz Pentium-III processors running under the Linux operating system.

EXAMPLE 3

All of the 1D shape signatures and 2D shape signatures (50,000 reflections per signature) generated in Example 1 were used as queries against the NCI shape signatures database generated in Example 2. The best 50 hits for each query were collected. Searches were carried out using 1D and 2D shape signatures, along with either L_1 or R_1 metrics (Equations. 1-2, and 3-4 respectively) for a total of four searches. Six query compounds, comprising a set that is both structurally diverse and biologically interesting, were selected for detailed examination.

A special concern when comparing a query against a large database is the distribution of scores. To be useful, a search method must exhibit a high degree of selectivity, so that truly interesting hits have scores that differ markedly from the mean. In other words, it should be possible to identify a reasonable cutoff score which can be applied to extract a relatively small and meaningful set of hits from a diverse target database. To examine the character of the scores distribution for shape signatures, a special version of the search program for the L_1 metric was prepared which accumulated score statistics in a file. It was thus possible to accumulate the distribution of scores from this relatively large database, and to express the number of observed hits as a function of score for each query molecule.

FIG. 16A lists top 1D hits for molecules from the Tripos fragment database used as queries against the NCI chemical library. The format of the table follows that of FIG. 14A, using the same queries and displaying results for 1D shape signatures, but in this table the score columns correspond to the use of different metrics L_1 and R_1, rather than having segment culling enabled or disabled. (Segment culling was used exclusively in preparation of the NCI shape signatures database, so there is no "non-culled" case to compare to). The hits are labeled by NCI compound IDs (CAS). The top five hit structures for each query are displayed in FIG. 16B.

We note at the outset that of the six queries shown, only 1,2,3,4-tetradihydroisoquinoline and adenine are in the NCI database subset used in these searches. In the case of 1,2,3,4-tetradihydroisoquinoline, the NCI entry (CAS #91-21-4) corresponding to the query is selected as the top hit, while for adenine the corresponding hit (CAS #73-24-5) appears in the third position in the hit list. We note in the latter case that the top hit (CAS #10325-61-8) differs from adenine only in the substitution of an amine nitrogen with an oxygen, leading to structures that are very similar in shape; furthermore the scores of #73-24-5 and #10325-61-8 differ by only 0.004 probability units. Given the probabilistic nature of the method and the presence of competing structures of almost identical shape, the "best" chemical structure will not always top the hit list. Also, small differences in the conformation of query and target compounds may influence the order of hits.

The other queries locate target compounds that are generally of very similar structure. Some interesting and initially unexpected hits: for query 1,2,3,4-tetradihydroisoquinoline, the hit #578-54-1 is seen to have the same structure as the query, but with the amine-containing ring opened; for query 5H-dibenz[b,f]azepin, hit #6279-16-9 is similar to the query, but with the central ring opened. This ability to locate "approximate" matches is an interesting feature of the shape signatures approach. It is also noted that hit #82-53-1 for query 5H-dibenz[b,f]azepin, clearly bears a weak resemblance to the query despite a low score. This false positive is to be expected as the discussed method collapses a large space of chemical structures onto a comparatively small descriptor space. Hits under the R_1 metric are similar to those under L_1, involving for some queries the introduction of new hits, but more often simply the reordering in ranking of existing hits.

In FIG. 17A there is presented the results for the six query compounds when used in 2D shape signature searches against the NCI database. Despite the much larger size of the NCI database compared to the Tripos fragment database, only three of the queries (1,2,3,4-tetrahydroisoquinoline, 5H-dibenz[b,f]azepin and adenine) have clearly significant hits (with top scores less than 0.1) and one other (1,4,6-gonatriene-3,17-drone) has hits of borderline significance (with scores all larger than 0.15). Structures for the three low-scoring queries are shown in FIG. 17B.

Examination of these hits, and comparison to FIG. 16B illustrates the role that electrostatics plays in selecting compounds. The hits for 1,2,3,4-tetrahydroisoquinoline all exhibit an electronegative nitrogen in a position identical or close to that found in the query. This is not the case for the 1D shape signature search where there is less consistency in the appearance of electronegative atoms in the hits. (We point out in both the 1D and 2D searches, the query compound appears as the top hit.) For 5H-dibenz[b,f]azepin, the top 2D hit has somewhat weaker shape similarity compared to the top 1D hit, including a cyclopropyl motif not found in the query; at the same time it includes an electronegative nitrogen at a position homologous to that of the query. The top hit is followed by hits that arguably exhibit weaker shape similarity to the query than some of the corresponding 1D hits, but which include an electronegative nitrogen at a position similar to the query. Finally, hit #5 for the 2D-MEP search is the compound found as hit #1 in the 1D search. For adenine, the 2D-MEP search produces hits which in every case contain nitrogens at positions homologous to the query. Compound #2846-89-1 is an interesting "substructure match." The top hit in this case is the query, while in the 1D search the query molecule appeared as the #3 hit.

Searching our NCI shape signature database (113,331 compounds) using a 1D shape signature query required on average 133 sec on a 450 MHz Pentium-III processor running the Linux operating system. This corresponds to 1.17 msec per comparison (which should be compared to the figure of 370 μsec quoted above for a 1.5 GHz machine). The average time per comparison for a 2D shape signature search was 3.7 msec.

For 1D searches, a distance of 0.05 or less in the metrics usually corresponded to strong shape similarity, while the range 0.05-0.1 is a borderline region where interesting hits may be mixed with "substructure" matches. For 2D-MEP searches under L_1 the corresponding ranges are 0-0.1 and 0.1-0.2. There should be a relatively small number of close hits for a given query. In contrast, the vast majority of target compounds should be unambiguously assigned as weak matches.

In FIG 10A there is shown the distributions of 1D scores for the three query compounds of FIG. 17B. FIG. 10B shows the distributions of 2D scores for these molecules using the L_1

metric. These two FIGS. include a logarithmic axis for the number of compounds observed for a given range of scores. Figs. 11A and 11B show the same distributions as FIGS 10A and 10B but with linear vertical axes. FIG 10 highlights the numbers of compounds observed at the extreme left of the distribution, where structurally-interesting matches are expected, while FIG. 11 provides a better sense of the shapes of the distributions, which appear to be locally Gaussian but with a significant shoulder. This gives the strong impression of being well-represented by a sum of two Gaussians, an observation that may be of practical significance for developing rapid tests for the significance of hits.

Given these distributions, we can directly compute the cutoff score needed to select a given percentage of-compounds in the NCI database. For a database of this size, two useful cutoffs are those needed to select 0.1 % and 0.01 % of the compounds, corresponding to approximately 100 and 10 compounds.

Noteworthy are the appearance of approximate and substructure matches as the scores increase. A general impression is that as scores increase, one observes first close matches, then hits that correspond to substructures or rearrangements of the query, and finally to hits that exhibit no clear similarity to the query. This allows identification of compounds that have at least partial similarity to an active compound, and which may be able to mimic a subset of the interactions of the query with a target receptor.

EXAMPLE 4

The HIV protease inhibitor Indinavir was used as a starting framework. As shown in FIG. 1, the compound includes pyridine (101), t-butyl forinamide (103), phenyl (105) and benzocyclopentanol (107) as substituents, which are attached to a framework containing piperazine, a peptide group, and a central hydroxyl which marks the site of the transition state analogue presented by the inhibitor.

Rather than attempt to find receptor-based matches to the entire binding site, the approach of finding matches to receptor subsites was used. Those subsites were defined by excising these substituents one-at-a-time from the experimental complex of the inhibitor and the native protease molecule. In this way four separate subsites were generated, each marked with a SYBYL dummy atom attached to the portion of the inhibitor that remained. One of these sites, (101), was largely exposed to solvent and did not provide a well-defined, enclosed pocket; it was omitted from the analysis below, and the original substituent (pyridine) was retained at this position.

Ray-tracing was performed in each pocket with 50,000 reflections, and the ray traces were used to generate 1D shape signatures using ray-trace segment length for the three sites considered (103), (105) and (107). These were used to search the NCI database for compounds of shape similar to the pocket volumes (or stated another way, of shape complementary to the receptor subsites). Parameters were identical to those used in the ligand-based searches discussed in the prior examples.

Once a collection of hits was assembled for each subsite, the hits were attached to the framework. This was done using a custom SYBYL application program called ALMS for Automated Ligand-binding with Multiple Substitutions. The program was written in the SYBYL programming language (SPL). A general flowchart of the program is shown in FIG. 12 in step (1201) each non-ring hydrogen of an NCI hit molecule was considered as a possible attachment point, in step (1203) each ring hydrogen as a possible attachment point through an added methylene carbon was considered. In this way, a single hit molecule was "exploded" into a family of fragments, each with a single free valence, marked by a dummy atom. In step (1205) a fragment is attached to the framework by removing the dummy atoms on both the hit and the target inhibitor site, and replacing these with a single bond linking the inhibitor and the fragment. The orientation of the newly-attached fragment was then optimized using FlexiDock, the genetic-algorithm-based optimizer included with the SYBYL modeling package. Default force-field settings were used in FlexiDock (including hydrogens with reduced van der Waals radius and epsilon parameter), and in all calculations the genetic algorithm proceeded for 500 generations.

Each framework variable site was considered individually, with additions of all the fragments generated from the best n hits for a particular site carried out with the substituents of the starting inhibitor in place at all the other sites. The FlexiDock inhibitor-receptor interaction energies computed after adding all of the fragments targeted to a particular site were used to rank the fragments for that site. Next, the top k, m and n fragments for sites (103), (105), and (107) respectively were added in all possible combinations, with precomputed optimized geometry, generating k*m*n ligand molecules. A final energy minimization was performed in the field of the frozen receptor for each ligand, followed by an updated computation of the interaction energy. The interaction energies so computed were used to rank the table of putative inhibitors finally generated.

FIG. 13A shows the receptor subsite created by removing the (103) substituent (t-butyl formamide) from the Indinavir framework, along with the associated ray-trace. FIG. 13B shows the corresponding results for the (105) substituent, and FIG. 13C shows the ray-trace for the (107) substituent.

The top fifty hits for each of the subsite queries were examined, and it was found that in each hit set there were many examples of closely-related structures (this was especially the case with the list for substituent (107)). Subsets of structures with high similarity were identified in each list, and only one representative compound from each set was retained in an effort to assemble a structurally-diverse group of NCI hit compounds for each subsite. This yielded 27 compounds for substituent (103), 40 for substituent (105) and 12 for substituent (107). After exploding each hit by assignment of all possible attachment points, there were a total of 377 fragments for substituent (103) 275 for substituent (105), and 108 for substituent (107). Each fragment was attached to its target inhibitor site and optimized as described above. The selection of NCI hit compounds implied a total of 11,196,900 possible inhibitor structures. FIG. 18 shows the best three fragments for each subsite, ranked by FlexiDock energy realized after attachment and optimization, along with the CAS ID number for the source NCI compound. We also show for comparison the substituent found at the same position in Indinavir.

To construct a collection of trial inhibitor structure, the best ten fragments for each of the three variable sites were selected, and inhibitors were constructed using all possible combinations of the selected fragments. Each selected fragment was attached to its target site with the FlexiDock-optimized conformation determined in the first phase of the procedure. This produced 1,000 initial structures. Interaction- and self-energies of all the compounds were computed by a utility in ALMS, which continued by ranking the compounds in order of ascending energy. The best fifty compounds were selected for energy minimization in the field of the frozen receptor.

The last phase of this "semi-automated" inhibitor design included generating a rough estimate of binding energy for each of the best fifty inhibitors. This involved removing the inhibitor (with optimized geometry) from the receptor and allowing it to minimize in isolation. Subtracting this minimized energy from the self-energy of the compound when docked provided an estimate of inhibitor strain energy, and this positive quantity was then added to the optimized inhibitor-receptor interaction energy to provide a binding energy estimate. Obviously this simple estimate did not take entropic factors into account, nor receptor flexibility.

All computations with SYBYL were carried out on a Silicon Graphics workstation. Total computing time can be divided into that required for the following phases: Attachment of 760 fragments to their respective inhibitor sites, followed by optimization using FlexiDock, 5.45 hr ; generation of 1,000 trial inhibitors, and initial computation of interaction energy, 1.49 hr; final minimization (1,000 steps max.) of best 50 inhibitors, 4.7 hr. Total CPU time was thus approximately 11.6 hr. The time required to scan the NCI database using the receptor-based shape signature queries was approximately 9 minutes (carried out on 550 MHz Pentium-III processors running Linux).

Three representative inhibitors proposed by this procedure are shown in FIG. 19, along with their estimated binding energies. Indinavir is included for comparison (its binding energy estimate was derived from the crystal structure, using the same protocol described above). Most of the top-ranking inhibitors involve combinations of a small selection of substituents at the three variable sites; all of the best-scoring compound had the same fragment (derived from compound #18650-61-8) at the substituent (107) position. Twenty-three of the inhibitors designed by our procedure have an estimated binding energy lower than -100 kcal/mole, and are predicted to be better binders than Indinavir.

While the invention has been disclosed in connection with certain preferred embodiments, this should not be taken as a limitation to all of the provided details. Modifications and variations of the described embodiments may be made without departing from the spirit and scope of the invention, and other embodiments should be understood to be encompassed in the present disclosure as would be understood by those of ordinary skill in the art.